CLAIMS:

1. An automated identification methodology for assembling document related hyperlinked pages comprising:

performing a page-level link analysis that identifies those hyperlinks on a page linking to a candidate document page potentially part of the document;

performing a recursive application of the page-level link analysis to the linked candidate document page and any further nested candidate document pages thereby identified, until a collective set of identified candidate document pages is assembled; and,

performing a document-level analysis that examines the collective set of identified candidate document pages for grouping into one or more documents.

2. The method of claim 1 wherein the page-level link analysis includes retrieval of referenced pages.

3. The method of claim 1 wherein the page-level link analysis includes examination of contextual clues.

4. The method of claim 3 wherein the contextual clue is a particular class of content item associated with the hyperlink.

5. The method of claim 4 wherein the class of content item is a class of text.

6. The method of claim 5 wherein the class of text is a directional word or phrase.

7. The method of claim 4 wherein the class of content item is a class of image.

8. The method of claim 7 wherein the class of image is an image containing a directional symbol.

9. The method of claim 4 wherein a textual clue is obtained for the image.

10. The method of claim 1 wherein the page-level link analysis includes the identification of progression links.

11. The method of claim 3 wherein the contextual clue is the presence of at least one other hyperlink nearby with the document description.

12. The method of claim 3 wherein the contextual clue is the similarity of the hyperlink destination to that of other hyperlinks with the document.

13. The method of claim 1 wherein the page-level link analysis includes the identification of tables of contents.

14. The method of claim 1 wherein the document-level analysis includes the identification of pages forming a chain of progression links.

15. The method of claim 1 wherein the document-level analysis includes identifying the pages listed in a table of contents.

16. The method of claim 1 wherein the document-level analysis includes identifying as part of the document the page containing the table of contents.

17. The method of claim 1 wherein the document-level analysis includes the similarity of candidate pages.

18. The method of claim 17 wherein the similarity includes the location at which the page is stored.

19. The method of claim 17 wherein the similarity includes the similarity of meta-data associated with the page.

20. The method of claim 19 wherein the meta-data includes the author identification.

21. The method of claim 17 wherein the similarity includes similar style specifications.

22. The method of claim 17 wherein the similarity includes similar page layout.

23. The method of claim 17 wherein the similarity includes similar logical structure of the page content.

24. The method of claim 17 wherein the similarity includes the presence of at least one similar content item on each page.

25. The method of claim 1 wherein the document-level analysis includes analysis of the topological structure of the linked pages.

26. A system identification methodology for assembling a hyperlinked document comprising:

performing a page-level link analysis that identifies those hyperlinks on a page linking to a candidate document page further comprising a methodology of:

identifying possible progression links, and;

identifying possible table of content links;

performing a recursive application of the page-level link analysis to the linked candidate document page and any further nested candidate document pages thereby identified, until a collective set of identified candidate document pages is assembled; and,

performing a document-level analysis that examines the collective set of identified candidate document pages for grouping into one or more documents.

27. The method of claim 26 wherein the page-level link analysis includes examination of contextual clues.

28. The method of claim 27 wherein the contextual clue is a particular class of content item associated with the hyperlink.

29. The method of claim 28 wherein the class of content item is a class of text.

30. The method of claim 29 wherein the class of text is a directional word or phrase.

31. The method of claim 28 wherein the class of content item is a class of image.

32. The method of claim 31 wherein the class of image is an image containing a directional symbol.

33. The method of claim 28 wherein a textual clue is obtained for the image.

34. The method of claim 27 wherein the contextual clue is the presence of at least one other hyperlink nearby with the document description.

35. The method of claim 27 wherein the contextual clue is the similarity of the hyperlink destination to that of other hyperlinks with the document.

36. The method of claim 26 wherein the document-level analysis includes the identification of pages forming a chain of progression links.

37. A system identification methodology for assembling a hyperlinked document comprising:

performing a page-level link analysis that identifies those hyperlinks on a page linking to a candidate document page further comprising a methodology of:

identifying possible progression links;

identifying possible table of content links, and;

examining the possible progression links and the possible table of content links for common characteristics;

performing a recursive application of the page-level link analysis to the linked candidate document page and any further nested candidate document pages thereby identified, until a collective set of identified candidate document pages is assembled; and,

performing a document-level analysis that examines the collective set of identified candidate document pages for grouping into one or more documents.

38. The method of claim 37 wherein the page-level link analysis includes examination of contextual clues.

39. The method of claim 38 wherein the contextual clue is a particular class of content item associated with the hyperlink.

40. The method of claim 39 wherein the class of content item is a class of text.

41. The method of claim 40 wherein the class of text is a directional word or phrase.

42. The method of claim 39 wherein the class of content item is a class of image.

43. The method of claim 42 wherein the class of image is an image containing a directional symbol.

44. The method of claim 39 wherein a textual clue is obtained for the image.

45. The method of claim 38 wherein the contextual clue is the presence of at least one other hyperlink nearby with the document description.

46. The method of claim 38 wherein the contextual clue is the similarity of the hyperlink destination to that of other hyperlinks with the document.

47. The method of claim 37 wherein the document-level analysis includes the identification of pages forming a chain of progression links.

48. The method of claim 37 wherein the document-level analysis includes the identification of pages linked to by the same tables of contents.